Facilitating Scientific Discovery in the Digital Age

Victoria Stodden
Department of Statistics
Columbia University

The Future of Scientific Knowledge Discovery in Open Networked Environments
BRDI/NAS Symposium and Workshop
Mar 10, 2011

Computation Emerging as Central to the Scientific Endeavor

For example, in statistics,

JASA June	Computational Articles	Code Publicly Available
1996	9 of 20	0%
2006	33 of 35	9%
2009	32 of 32	16%

A Crisis in Computational Science

- Computational methods becoming central to the scientific enterprise:
  - enormous, and increasing, amounts of data collection,
  - intellectual contributions now encoded in software,
  - typical scientific results rely on both data and code.
- Data and code typically not made available, rendering published results unverifiable, not reproducible.

➡ A Credibility Crisis

Reproducibility is Central to the Scientific Method

- Other branches of science incorporate reproducibility of results:
  - deductive branch (mathematics, formal logic): the well-defined concept of the proof,
  - inductive branch (experimental sciences): machinery of hypothesis testing, structured communication of methods and protocols.
- Computational Science must develop standards for reproducibility before it can be considered a third branch of the scientific method,

➡ Data and Code Sharing, with publication.

Framing Principle for Scientific Communication: *Reproducibility*

Data and code sharing *at the time of publication* is *imperative*:

- computational, data-driven, science must be *reproducible*,
- code and data contain the methodology,
- all but impossible to replicate published computational results without access to the underlying code and data,
- consequences for verifiability (ClimateGate, Duke Clinical Trials...) and public confidence in science.

What's missing?

- talks so far emphasize how science itself hasn't changed (Brahe/Kepler), but the scale, scope, and nature of the research has.

➡ different skills (Hey) and verifiability (Friend) needed,

➡ infrastructure, incentives must adapt:

  ➡ tool development for reproducibility and collaboration,

  ➡ openness in the publication of scientific discoveries.

Tool Development

- workflow tracking and provenance ie. Vistrails.org and many others,
- automatic cloud repository and unique identifiers for published results (Donoho, Gavish 2011),
- collaborative tools ie. colwiz,
- versioning and facilitation of collaboration.

Publication and Peer-Review

- today, code and data and not generally published or shared,
- code and data not typically subject to review, or even made openly available,
- ... yet it is a crucial part of the methodology need for replication.

Journal Policy

- Different approaches by journals:
  - may offer unreviewed "supplemental materials" section,
  - may require data and/or code to be provided upon request (Science as of Feb 11 2011),
  - may employ an Associate Editor for Reproducibility (Biostatistics, Biometrical Journal) or replicate results (ACM SIGMOD),
  - may publish correspondence from the review process (Molecular Systems Biology, The European Molecular Biology Organization Journal),
  - new journals, ie. Open Research Computation, BMC Data Notes
  - ignore the issue..

Funder Policy

- NIH PubMed Central, Open Access (idea: PubCentral),
- NSF peer-reviewed Data Management plan (Jan 13, 2011),
- NSF/OCI report on Virtual Communities (Dec, 2010),
- IOM "Review of Omics-based Tests for Predicting Patient Outcomes in Clinical Trials,"
- hesitation to fund software or infrastructure such as repositories (examples),
- idea: fund pilot projects that are reproducible.

Incentives and Open Questions: Citation and Contributions

- Collaborative efforts in database building?
  - differential citation? (web vs article citation, microcitation)
  - database versioning (e.g. King and Altman 2007, Donoho and Gavish 2011)
  - citizen contributions? (Galaxy Zoo, Open Dinosaur Project)
- Code development? review?
- Code maintenance for reproducibility, scientific reuse?
  - platform building (DANSE, Wavelab, Sparselab)
  - open source software as a model?

Barriers to Data and Code Sharing in Computational Science

Survey of Machine Learning Community (Stodden, 2010):

Code		Data
77%	Time to document and clean up	54%
52%	Dealing with questions from users	34%
44%	Not receiving attribution	42%
40%	Possibility of patents	-
34%	Legal Barriers (ie. copyright)	41%
-	Time to verify release with admin	38%
30%	Potential loss of future publications	35%
30%	Competitors may get an advantage	33%
20%	Web/disk space limitations	29%

Groundswell within the Computational Sciences

Previously:

- AAAS 2011 Symposium on "The Digitization of Science: Reproducibility and Interdisciplinary Knowledge Transfer"
- SIAM CSE 2011 Minisymposium on "Verifiable, Reproducible Computational Science"
- Yale Roundtable on Data and Code Sharing in the Computational Sciences 2009
- ACM SIGMOD conferences

Groundswell..

Upcoming:

- AMP 2011 "Reproducible Research: Tools and Strategies for Scientific Computing"
- AMP / ICIAM 2011 "Community Forum on Reproducible Research Policies"

Challenges to Open Science

- “Taleb Effect” - scientific discoveries as (misused) black boxes,
- nefarious uses?
- black boxes and opacity in software (why the traditional methods section is inadequate, massive codebases),
- lock-in: calcification of ideas in software?
- independent replication discouraged?
- policy maker engagement: finding support for our norms,
- Commercial incentives for the scientist/university (Bayh-Dole).

References

- “Enabling Reproducible Research: Open Licensing for Scientific Innovation”
- “The Scientific Method in Practice: Reproducibility in the Computational Sciences”
- “Open Science: Policy Implications for the Evolving Phenomenon of User-led Scientific Innovation”

available at http://www.stanford.edu/~vcs